

DOCUMENT RESUME

ED 435 979

CS 013 756

AUTHOR Wright, Benjamin D.; Stenner, A. Jackson
TITLE Readability and Reading Ability.
PUB DATE 1998-06-00
NOTE 28p.; Paper presented to the Australian Council on Education Research (June 1998).
PUB TYPE Guides - Non-Classroom (055) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability Identification; Higher Education; *Measurement Techniques; *Readability Formulas; *Reading Ability; Reading Achievement; Reading Comprehension; *Reading Diagnosis
IDENTIFIERS *Lexile Scale of Reading; *Reading Appraisal Guide

ABSTRACT

This document discusses the measurement of reading ability and the readability of books by application of the Lexile framework. It begins by stating the importance of uniform measures. It then discusses the history of reading ability testing, based on the assumption that no researcher has been able to measure more than one kind of reading ability. The concept of "lexiles" is introduced as a measure for readability which uses word familiarity and sentence length. The Lexile formula is explained, and processes for using the Lexile concept and applying it to reading instruction are discussed at length. The paper ends by relating Lexile scores to jobs, amount of education, and income. Contains 44 references, 2 notes, and 5 tables and 20 figures of data. (EF)

Readability and Reading Ability

June, 1998 Presentation to Australian Council on Education Research (ACER)

by Benjamin D. Wright, with A. Jackson Stenner

Uniform Measures

The world of education has long been waiting for a sunrise. Believe it or not, a popular compilation of educational tests lists 97 different reading tests (Mitchell, 1985). This situation produces 97 different "reading ability measures." What a mess! But now, with the dawn of uniform educational measures, the sun is rising here in Melbourne.

When I was a physicist, I came to appreciate the essential part uniform measures play in science. In the 17th century, there were many ways to observe the effects of heat. It was thought, therefore, that there were many kinds of heat. That was a brutal barrier to progress. Nevertheless, 17th century scientists thought they were observing "57 varieties." After all, is not bathtub heat different from teacup heat, different from cauldron heat, different from fireplace heat — all of which are different from the heat of the sun? Eventually, it was discovered that it was not only desirable but also necessary to have just one kind of heat. Today, for science and commerce, we do our thinking about heat in terms of one entirely abstract unit, the "degree." Whether it's a Kelvin, Celsius, or Fahrenheit degree does not matter. We know exactly how to get from one to another. They all measure, what we insist is, the same one kind of heat.

Measures are older than talking. Birds measure. So do bees. Our own measures evolved from our bodies — our feet, our arms, our hands, our fingers. An inch is the distance from thumb tip to knuckle. A span is the distance between thumb tip and little finger. A cubit is the length of a forearm. A fathom is the distance between outstretched arms. A pace is two steps. A furlong is 200 paces. A mile 1,000.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

2

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL HAS
BEEN GRANTED BY

B D Wright

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

1

Abstractly equal units of length were counted on before the oldest fragments of writing. Figure 1 is Moses' plan for the Tabernacle.

Without approximations to equal units, Babylonians, Egyptians and Hebrews could not have imagine, let alone built, their towers.

Fair measurement is embedded in Judeo-Christian morality. But the "perfect and just measure" demanded in Deuteronomy 25 is an ideal that can only be approximated in practice. The "weight" referred

to is a shekel stone, understood to weigh 11.4 ounces. However, archeologists have never found two shekel stones that weighed exactly the same. No technology, no matter how advanced, can fabricate perfect weights. Nevertheless, even when Deuteronomy was written, we already understood the essential necessity and justice of fair units.

The necessity of uniformity in the representation of quantity appears again in King John's Magna Carta. Without the ideal of uniform measures, there would be no

money. There would be no fitted clothes, because there would be no way to fit them. Imagine what life would be like if there were no abstract unit of length like the inch. Suppose that taking an inch were complex — differing with every situation and material. Imagine that wood inches were different from brick inches, were different from steel inches. We would not have civilization. We would have a mess — a mess like the mess that permeates most of what we misleadingly refer to as "educational tests and measurements."

Figure 1
Exodus 26

2. THE LENGTH OF ONE CURTAIN SHALL BE EIGHT AND TWENTY CUBITS , AND THE BREADTH OF ONE CURTAIN FOUR CUBITS ; AND EVERY ONE OF THE CURTAINS SHALL HAVE ONE MEASURE .

Figure 2
Deuteronomy 25

13. THOU SHALT NOT HAVE IN THY BAG DIVERSE WEIGHTS , A GREAT AND A SMALL .
14. THOU SHALT NOT HAVE IN THINE HOUSE DIVERSE MEASURES , A GREAT AND A SMALL .
15. THOU SHALT HAVE A PERFECT AND JUST WEIGHT , A PERFECT AND JUST MEASURE .

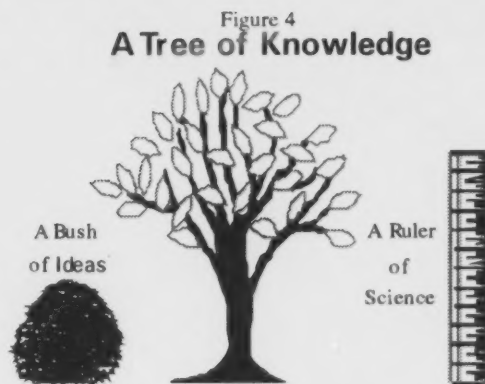
Figure 3
The Magna Carta

35. THERE IS TO BE ONE MEASURE OF WINE AND ALE AND CORN WITHIN THE REALM , NAMELY THE LONDON QUARTER , AND ONE BREADTH OF CLOTH , AND IT SHALL BE THE SAME WITH WEIGHTS .

The Evolution of Science

The study of any subject begins with tangles of speculations. Ideas branch in all directions. But as we work through the tangle, we connect what we experience with what we see. We coax our ideas into shape, form unities, develop lines of inquiry. We fit our ideas together and

make them something. We evolve our bush of ideas into a tree of knowledge. The bush was a tangle. The tree has direction. Our final step in wrestling an useful abstract assertion from a complex concrete confusion is to carve a ruler out of our tree. The ruler does not exist until we imagine it and carve it. The carving is not perfect. It is just an approximation. But what it approximates — a perfectly straight line — enables us to use it as though it were marked off in perfectly equal intervals.



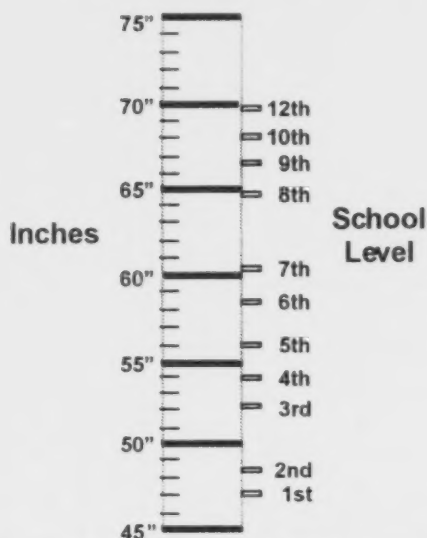
We can pace off land in somewhat equal steps. But steps inevitably vary according to conditions. To produce reliable measurements, we need something more reproducible than pacing. The scientific measurement of length was born as we connected our experience of stride with man made marks on straight pieces of wood extracted from tree trunks. A piece of tree is more stable than anyone's paces. A ruler does not change its bench marks. When we grow a confusing bush of tangled ideas into a tree of useful knowledge and make a ruler, then we can plan and build a pyramid, a temple, a house — and also measure the height of a child (Rasch, 1980).

The Imaginary Inch

An inch is *pure*, abstract and without content. It has no meaning of its own. It is an imaginary unit of length. A height of inches, however, has meaning. As we grow, we learn the advantages to growing taller. Brick size has meaning. As we build, we learn the advantages of same-sized bricks. What makes bricks useful is that their interchangeability is maintained by approximations to the fiction we call an inch.

It is essential that our idea of an abstract inch is always the same. If we let our idea of an inch change each time we made a measure, we could not produce useful bricks or keep track of our child's growth. As our child grew, we would not know by how much they had grown. But with an uniform unit of measurement, like an inch, when we measure the height of our children, we can refer to last year — or perhaps to the height of an average second grader because, as it turns out, child height is related to school grade. Figure 5 shows

Figure 5
**Educational Status
by Average Height**

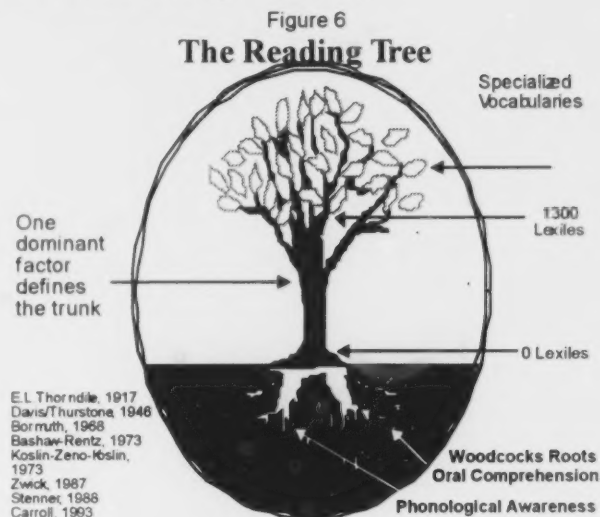


how we can guess what grade a child is in by how tall they are — and also how tall they are by what grade they are in. That is an understanding based entirely on applications of rulers. The applications would be useless without that single, unvarying inch that our rulers approximate.

No metric has content of its own. The ruler, with its equal measurement units is merely an approximate realization of a pure idea — an ideal which we invented from tangled experiences of length — invented to make uniform measures available for any application we may care to undertake.

One Kind of Reading Ability

Let's turn to the measure of reading. We can think of reading as the tree in Figure 6. It has roots like oral comprehension and phonological awareness. As reading ability grows, a trunk extends through grade school, high school and college branching at the top into specialized vocabularies. That single trunk is longer than many realize. It grows quite straight and singular from first grade through college.



Reading has always been the most researched topic in education (Thorndike, 1965). There have been many studies of reading ability, large and small, local and national. When the results of these studies are reviewed, one clear picture emerges. Despite the 97 ways to test reading ability (Mitchell, 1985), many decades of empirical data document definitively that no researcher has been able to measure more than one kind of reading ability. This has proven true in spite of intense interest in discovering diversity. Consider three examples: the 1940's Davis Study, the 1970's Anchor Study and six 1980's and 1990's ETS studies.

Davis – 1940's

Fred Davis went to a great deal of trouble to define and operationalize nine kinds of reading ability (1944). He made up nine different reading tests to prove the separate identities of his nine kinds. He gave his nine tests to hundreds of students, analyzed their responses to prove his thesis, and reported that he had established nine kinds of reading. But when Louis Thurstone reanalyzed Davis' data (1946), Thurstone showed conclusively that Davis had no evidence of more than one dimension of reading.

Anchor Study – 1970's

In the 1970's, worry about national literacy moved the U.S. government to finance a national Anchor Study (Jaeger, 1973). 14 different reading tests were administered to a great many children in order to uncover the relationships among the 14 different test scores. Millions of dollars were spent. Thousands of responses were analyzed. The final report required 15,000 pages in 30 volumes — just the kind of document one reads overnight, takes to school the next day and applies to teaching (Loret et al, 1974). In reaction to this futility, and against a great deal of proprietary resistance, Bashaw and Rentz were able to obtain a small grant to reanalyze the Anchor Study data (1975, 1977). By applying new methods for constructing objective measurement (Rasch, 1980; Wright and Stone, 1979), Bashaw and Rentz were able to show that all 14 tests used in the Anchor Study — with all their different kinds of items, item authors, and publishers — could all be calibrated onto one linear “National Reference Scale” of reading ability.

The essence of the Bashaw and Rentz results can be summarized on one easy-to-read page (1977) — a bit more useful than 15,000 pages. Their one page summary shows how every raw score from the 14 Anchor Study reading tests can be equated to one linear National Reference Scale. Their page also shows that the scores of all 14 tests can be understood as measuring the same kind of reading on one common scale. The Bashaw and Rentz National Reference Scale is additional evidence that, so far, no more than one kind of reading ability has ever been measured. Unfortunately, their work had little effect on the course of U.S. education. The experts went right on claiming there must be more than one kind of reading — and sending teachers confusing messages as to what they were supposed to teach and how to do it.

ETS Studies – 1980's and 1990's

In the 1980's and 1990's, the Educational Testing Service (ETS) did a series of studies for the U.S. government. ETS (1990) insisted on three kinds of reading: prose reading, document reading and quantitative reading. They built a separate test to measure each of these three kinds of reading — greatly increasing costs. Versions of these tests were administered to samples of school children, prisoners, young adults, mature adults, and

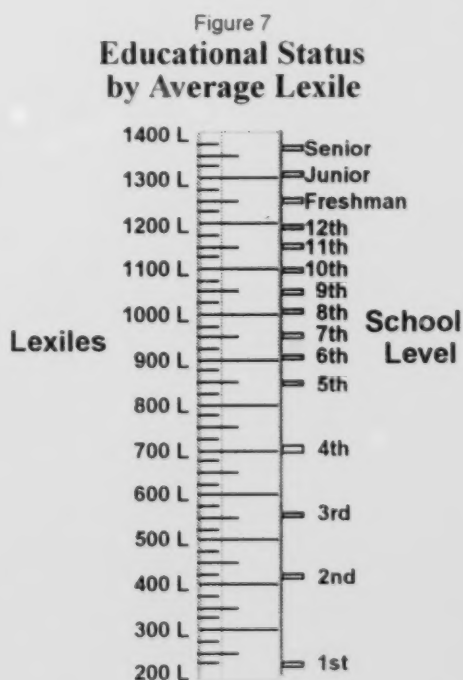
senior citizens. ETS reported three reading measures for each person and claimed to have measured three kinds of reading (Kirsch & Jungeblut, 1986). But reviewers noted that, no matter which kind of reading was chosen, there were no differences in the results (Kirsch & Jungeblut, 1993, 1994; Reder, 1996; Salganik & Tal, 1989; Zwick, 1987). When the relationships among reading and age and ethnicity were analyzed, whether for prose, document, or quantitative reading, all conclusions came out the same.

Later, when the various sets of ETS data were reanalyzed by independent researchers, no evidence for three kinds of reading measures could be found (Bernstein, & Teng, 1989; Reder, Rock and Yamamoto, 1994; 1996; Salganik and Tal, 1989; Zwick, 1987). The correlations among ETS prose, document and quantitative reading measures ranged from 0.89 to 0.96. Thus, once again and in spite of strong proprietary and theoretical interests in proving otherwise, nobody had succeeded in measuring more than one kind of reading ability.

Lexiles

Figure 7 is a reading ruler. Its Lexile units work just like the inches. The Lexile ruler is built out of readability theory, school practice, and educational science. The Lexile scale is an interval scale. It comes from a theoretical specification of a readability unit that corresponds to the empirical calibrations of reading test items (Rasch, 1980; Stenner, 1997). It is a readability ruler. And it is a reading ability ruler.

Readability formulas are built out of abstract characteristics of language. No attempt is made to identify what a word or sentence means. The idea is not new. The Athenian Bar Association used readability



calculations to teach lawyers to write briefs in 400 B.C. (Chall, 1988; Zakaluk and Samuels, 1988). According to the Athenians, the ability to read a passage was not the ability to interpret what the passage was about. The ability to read was just the ability to read. Talmudic teachers who wanted to regularize their students studies, used readability measures to divide the Torah readings into equal portions of reading difficulty in 700 A.D. (Lorge, 1944). Like the Athenians, their concern in doing this was not with what a particular Torah passage was about, but rather the extent to which passage readability burdened readers.

In the twentieth century, every imaginable structural characteristic of a passage has been tested as a potential source for a readability measure: the number of letters and syllables in a word; the number of sentences in a passage; sentence length; balances between pronouns and nouns, verbs and prepositions (Stenner, 1997). The Lexile readability measure uses word familiarity and sentence length.

Lexile Accuracies

Table 1 lists the correlations between readability measures from the ten most studied readability equations and student responses to different types of reading test items. The columns of Table 1 report on five item types:

Lexile Slices;
SRA Passages;
Battery Test Sentences;
Mastery Test Cloze Gaps;
Peabody Test Pictures.

Table 1
Correlations between
Empirical & Theoretical
Item Dif/culties

Ten Readability Equations	Five Test Item Types				
	Lexile Slice	SRA Passage	Battery Sentence	Mastery Cloze	Peabody Picture
Lexile	.90	.92	.85	.74	.94
Flesch	.85	.94	.85	.70	.85
ARI	.85	.93	.85	.71	.85
FOG	.85	.92	.73	.75	.85
Powers	.82	.93	.83	.65	.74
Holquist	.81	.91	.81	.84	.86
Flesch-1	.79	.92	.81	.61	.69
Flesch-2	.75	.87	.70	.52	.71
Coleman	.74	.87	.75	.75	.83
Dale-Chall	.76	.92	.82	.73	.67

Adapted from Stenner, 1997

The item types span the range of reading comprehension items. The numbers in the table show the correlations between theoretical readability measures of item text and empirical item calibrations calculated from students' test responses. Consider the top row. The Lexile readability equation predicted how difficult Lexile slices would be for persons

taking a Lexile reading test at a correlation of 0.90, the SRA passage at 0.92, the Battery Sentence at 0.85, the Mastery Cloze at 0.74 and the Peabody Picture at 0.94 (Stenner, 1996, 1997). With the exception of the cloze items, these predictions are nearly perfect. Also note that the simple Lexile equation, based only on word familiarity and sentence length, predicts empirical item responses as well as any other readability equation — no matter how complex. Table 1 documents, yet again that one, and only one, kind of reading is measured by these reading tests. Were that not so, the array of nearly perfect correlations could not occur. Table 1 also shows that we can have a useful measurement of text readability and reader reading ability on a single reading ruler!

An important tool in reading education is the basal reader. The teaching sequence of basal readers records generations of practical experience with text readability and its bearing on student reading ability. Table 2 lists the correlations between Lexile Readability and Basal Reader Order for the eleven basal readers most used in the United States. Each series is built to mark out successive units of increasing reading

Table 2
Correlations between
Basal Reader Order & Lexile Readability

Basal Reader Series	Basal Units	r	R	R'
Ginn	53	.93	.98	1.00
HB JEagle	70	.93	.98	1.00
SFFocus	92	.84	.99	1.00
Riverside	67	.87	.97	1.00
HM (1983)	33	.88	.96	.99
Economy	67	.86	.96	.99
SFAmer Trad	88	.85	.97	.99
HB JOdyssey	38	.79	.97	.99
Holt	54	.87	.96	.98
HM (1986)	46	.81	.95	.97
Open Court	52	.54	.94	.97

Adapted from Stenner,
1997

r = raw R = corrected for attenuation R' = corrected for attenuation and range restriction

difficulty. Ginn has 53 units — from book 1 at the easiest to book 53 at the hardest. HBJ Eagle has 70 units. Teachers work their students through these series from start to finish. Table 2 shows that the correlations between Lexile measures of the texts of these basal readers and their sequential positions from easy to hard are extraordinarily high. In fact, when corrected for attenuation and range restriction, these correlations become approach perfection (Stenner, et al, 1992, 1996, 1997, 1998).

Each designer of a basal reader series used their own ideas, consultants and theory to decide what was easy and what was hard. Nevertheless, when the texts of these basal units are Lexiled, these Lexiles predict exactly where each book stands on its own reading

ladder — more evidence that, despite differences among publishers and authors, all units end up bench marking the same single dimension of reading ability.

Finally there are the ubiquitous reading ability tests administered annually to assess every student's reading ability. Table 3 shows how well theoretical item text Lexiles predict actual readers' test performances on eight of the most popular reading tests.

Table 3
Correlations between
Passage Lexiles & Item Readabilities

Tests	Passages Analyzed	Item Type	r	R	R'
SRA	48	Passage	.95	.97	1.00
CAT - E	74	Passage	.91	.95	.98
CAT - C	43	Passage	.83	.93	.96
CTBS	50	Passage	.74	.92	.95
NAEP	70	Passage	.65	.92	.94
Lexile	262	Slice	.93	.95	.97
PIAT	66	Picture	.93	.94	.97
Mastery	85	Cloze	.74	.75	.77

Adapted from Stenner, 1997

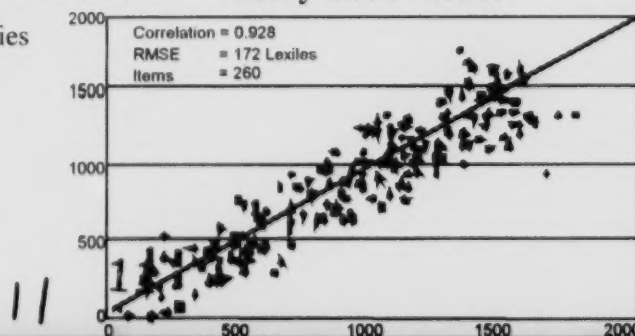
r = raw R = corrected for attenuation R' = corrected for attenuation and range restriction

The second column shows how many passages from each test were Lexiled. The third column lists the item type. Once again there is a very high correlation between the difficulty of these items as calculated by the entirely abstract Lexile specification equation and the live data produced by students answering these items on reading tests. When we correct for attenuation and range restriction, the correlations are just about perfect. Only the Mastery Cloze test, well-known to be idiosyncratic, fails to conform fully (Stenner, 1997).

What does this mean? Not only is only one reading ability being measured by all of these reading comprehension tests, but we can replace all the expensive data used to calibrate these tests empirically with one formula — the abstract Lexile specification equation. We can calculate the reading difficulty of test items by Lexiling their text without administering them to a single student!

Figure 8 puts the relationship between theoretical Lexiles and observed item difficulties into perspective. The uncorrected correlation of 0.93, when

Figure 8
Theory into Practice



disattenuated for error and corrected for range restrictions, approaches 1.00. The Lexile equation produces an almost perfect correlation between theory and practice (Stenner, 1997).

Figure 8 shows the extent to which idiosyncratic variations in student responses and item response options enter the process. Where does this variation come from? Item response options have to compete with each other or they do not work. But there has to be one correct answer. Irregularity in the composition of multiple choice options, even when they are reduced to choosing one word to fill a blank, is unavoidable. What the item writer chooses to ask about a passage and the options they offer the test taker to choose among are not only about reading ability. They are also about personal differences among test writers.

There are also variations among test takers in alertness and motivation that disturb their performances. In view of these unavoidable contingencies, it is surprising that the correlation between Lexile theory and actual practice is so high.

How does this affect the measurement of reading ability? The root mean square measurement error for a one item test would be about 172 Lexiles (Stenner, 1997). What are the implications of that much error? The distance from First Grade school books to Second Grade school books is 200 Lexiles. So we would undoubtedly be uneasy with measurement errors as large as 172 Lexiles. However, when we combine the responses to a test of 25 Lexile items, the measurement error drops to 35 Lexiles. And when we use a test of 50 Lexile items, the measurement error drops to 25 Lexiles — one eighth of the 200 Lexile difference between First and Second Grade books. Thus, when we combine a few of Lexile items into a test, we get a measure of where a reader is on the Lexile reading ability ruler, precise enough for all practical purposes. We do not plumb their depths of understanding. But we do measure their reading ability.

Lexile Items

One might now ask, how hard is it to write a Lexile test item? Figure 9 describes a study to find out whether Lexile items written by different authors produce usefully equivalent results (Stenner, 1998). Five apprentice item authors were each asked to choose their own text passages and to write their own response illustrated missing word options (Figure 10). Each author wrote 60 items spanning 900 to 1300 Lexiles. From these ($5 \times 60 = 300$) items, five 60 item tests were constructed

by drawing 12 items at random from each author. Then seven grade school students were given a different test each day for five days. This produced five measures for each student over the five days. And, by pooling days, five measures for each student over the five authors.

The question becomes, "Is the variation by author in a student's reading ability measure any larger than the variation by day?" If not, that would imply that writing useful Lexile test items, as in Figure 10, was not a problem, since even apprentice authors can do it well enough to obtain measures as stable as the differences in a person's reading performance from day to day.

Figure 9
Stability Study

**5 different item authors compose
5 different sets of 60 Lexile items
evenly sequenced from 900 to 1300 Lexiles**

**5 different 60 item tests are assembled
Each test containing
12 items selected at random
from each author's set of 60**

**7 grade school students take
a different 60 item test each day for 5 days**

**This produces, for each student
5 measures across 5 days balanced over authors
and
5 measures across 5 authors balanced over days**

Figure 10
An 800 Lexile Slice Test Item

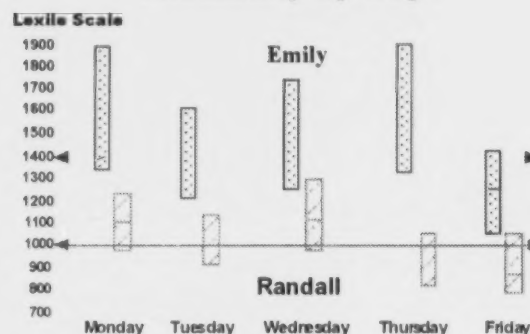
Wilber likes Charlotte better and better each day. Her campaign against insects seemed sensible and useful. Hardly anybody around the farm had a good word to say for a fly. Flies spent their time pestering others. The cows hated them. The horses hated them. The sheep loathed them. Mr. and Mrs. Zuckerman were always complaining about them, and putting up screens. Everyone _____ about them.

- a) agreed
- b) gathered
- c) laughed
- d) learned

from Charlotte's Web by E. B. White, 1952, New York: Harper and Row.

We know that each person's reading performance varies from day to day. Each performance depends on what is happening in our lives, what we have for breakfast, what happens at home, what happens at school, and how we feel about the test. Figure 11 shows the day to day results for

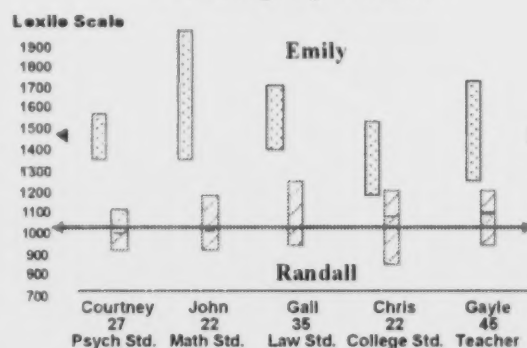
Figure 11
**Reading Ability
Instability by Day**



Emily and Randall. The vertical bars mark a 75 percent confidence region for the reading ability measure on each day. The up and down movements of the bars show how much these estimates of reading ability changed from day to day. On Monday Randall and Emily did relatively well. On Tuesday their performances sank. On Wednesday they came back. On Thursday Emily went up, but Randall went down. Finally, on Friday, they both went down. Figure 11 shows the differences a day makes in the reading performance of these two students. It reminds us that, when we talk about reading ability, we must remember that performances vary from day to day.

Figure 12 shows the variation in reading measures by item author. Notice that the variation among item authors in Figure 12 is no greater than the variation over days in Figure 11. No more noise is introduced into the Lexile way of making a reading measure by a difference among item authors than by the difference a day makes.

Figure 12
**Reading Ability
Stability by Author**



These five Lexile item authors were not experts. They were just well educated persons, instructed in Lexile item writing for four hours. Courtney, 27, is a psychology student.

John, 23, is a math student. Gail, 35, is law student. Chris, 22, is a football player. Gayle, 45, is a teacher.

Calculating Lexiles

Lexile measures of reading are easy to understand and easy to use.

Lexile readability — measured by word familiarity and sentence length — establishes how difficult a

READABILITY is passage reading difficulty.

READING ABILITY is ability to read passages.

Lexile reading ability is measured by finding out what Lexile passage readability a person can read with 75 percent success.

Success is defined as recognizing what words are needed to mend gaps inserted in passages.

text is to read. Lexile reading ability — measured by how well a reader is able to recognize words and connect them into sentences — establishes how able a reader is to read a text (Stenner, 1982, 1983, 1987).

The Lexile formula is based on two axioms.

The semantic axiom: the more familiar the words, the easier the passage is to read;

the more unfamiliar the words, the harder.

The syntactic axiom: the shorter the sentences, the easier the passage is to read;

the longer the sentences, the harder.

These axioms apply to whatever is read, quite apart from content. They apply whether we like what we are reading or not, whether it is prose, document or quantitative.

The Lexile system calculates passage readability from just these two characteristics — both of which are explicit in the passage. Sentence lengths are there to see. We count and average them. Word familiarities are obtained from compilations of word usage. The Lexile Analyzer uses John Carroll's sample of 5 million words (Word Frequency Book, 1971).¹

¹ The familiarity of the words used in a passage can be estimated from any comprehensive word usage compilation — A Basic Vocabulary of Elementary School Children, Henry D. Rinsland, 1945; The Teacher's Word Book of 30,000 Words, Edward L. Thorndike and Irving Lorge, 1952; The Word Frequency Book, John B. Carroll, Peter Davies and Barry

If readers do not know the words, they cannot read the passage. If they do know the words, they can begin to make the passage take shape by stringing its words into sentences. If they can make the sentences, they can read the passage and then, and only then, begin to think about what the passage has to say. Knowing the words and making the sentences sets the threshold for reading (Hitch, & Baddeley, 1974; Lieberman, et al, 1982; Shankweiler, & Crain, 1986; Miller & Gildea, 1987).

To Lexile a passage, we look up the occurrence frequency of each word. The Lexile Analyzer uses the average log word frequency and the logarithm of average sentence length. The final Lexile measure for the passage is a weighted sum of these two logarithms. Figure 13

Figure 13

How to Lexile a Book

Divide the BOOK into natural SLICES of 125-140 words.
For each TEXT SLICE i , determine:
Log mean sentence length = SL_i
Mean log word frequency = WF_i
Then
CALIBRATE SLICE i
at
 $Readability = 582 + 1768 * SL_i - 386 * WF_i$ Lexiles
The Lexile Measure of a Book
is Equal to the Lexile Level of a Reader
who Succeeds on 75% of that Book's Slices

shows how to Lexile a book. The coefficients in the formula are set to provide the most efficient balance between log word familiarity and log sentence length and to define a metric that reaches 1000 Lexiles from the books used in First Grade at 200 Lexiles to the books used in Twelfth Grade at 1200 Lexiles. The full Lexile range of readability goes from zero to 1800. The equation is simple. Word familiarity and sentence length are all there is to it. Figure 14 shows how to Lexile a reader.

Figure 14

How to Lexile a Reader

Test the READER
with L Response Illustrated
Lexile Calibrated SLICES
of Average slice Lexile = H
and slice Lexile Standard Deviation = S
Then
Count the READER'S right answers for Score = R
This Reader's Lexile MEASURE is
 $Reading Ability = H + (180 + S^2/1040) \log [R/L-R]$
The Lexile Measure of a Reader
is Equal to the Lexile Level of a Text for
which the Reader Succeeds on 75% of the Slices

Richman, 1971; The Educators' Word Frequency Guide, Susan M. Zeno, Stephen H. Ivens, Robert T. Millard and Raj Davvuri, 1995.

Lexile Relationships

Tables 4 and 5 illustrate some useful Lexile relationships. When a reader with a Lexile ability of 1000L is given a 1000L text, we expect them to experience a 75 percent success rate (Stenner, 1992). If the same reader is given a 750L text, then we expect their success rate to improve to 90 percent. If a text is at 500L, their success rate should improve to 96 percent. The more readers' Lexile reading abilities surpass the Lexile readability of a text, the higher their expected success rates. However, the more a text Lexile readability surpasses readers' Lexile reading abilities, the lower their expected success rates.

Table 4
**Success Rates for Readers
of Similar Ability with
Texts of Different Readability**

Reader Ability Lexile	Text Readability Lexile	Text Titles	Expected Success
1000	500	<i>Are You There God?</i>	96%
		<i>Its Me Margaret,</i> Blume	
1000	750	<i>The Martian Chronicles,</i> Bradbury	90%
1000	1000	<i>The Readers'</i> <i>Digest</i>	75%
1000	1250	<i>The Call of the Wild, London</i>	50%
1000	1500	<i>On Equality</i> <i>Among Mankind,</i> Rousseau	25%

Table 5
**Success Rates for Readers
of Different Ability with
Texts of Similar Readability**

Reader Ability Lexile	Sports Illustrated Readability Lexile	Expected Success
500	1000	25%
750	1000	50%
1000	1000	75%
1250	1000	90%
1500	1000	96%

Success rates are relative. They are the results of Lexile differences between readers and texts. The 250L difference between a 750L text and a 1000L reader results in the same success rate as the 250L difference between a 1000L text and a 1250L reader. Each reader-text combination produces 90 percent reading success. Success rates are centered at 75 percent because readers forced to read at 50 percent success report frustration, while readers reading at 75 percent report comfort, confidence and interest.²

² Squires, Huitt and Segars (1983) found that reading achievement for second-graders peaked when their success rate reached 75 percent. A 75 percent rate is also supported by the findings of Crawford, King, Brophy, and Evertson (1975).

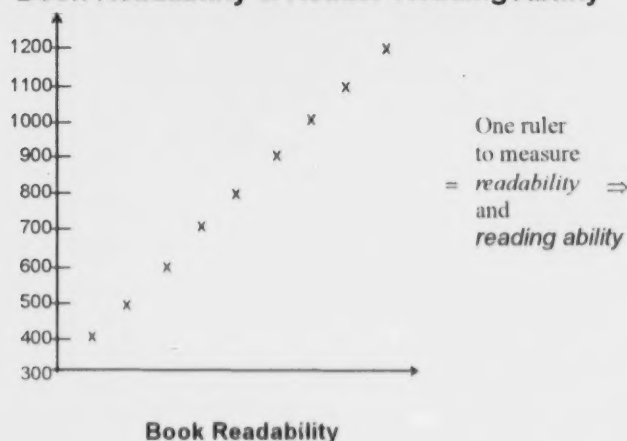
Each reader has their own range of reading comfort. As a result, there is a natural range of text readability that most motivates each reader to improve their reading ability. Some readers are challenged by a success rate as low as 60 percent. Others find that burdensome. Once a reader places themselves and their books in the Lexile Framework, they can discover what Lexile difference between their reading ability and text readability challenges them in the most productive way.

Book readability varies from page to page. Some books have a narrow range, their passages cluster around a common level. As we read these books, the reading challenge stays level. There are no hills or valleys. Other books have a wide range of readability. There are easy passages and hard passages. These books can enable us to use the momentum we gain from the easier passages to surmount the challenge of the harder ones. Overcoming this kind of resistance improves reading ability.

When we want to help a student read, we can Lexile them and then offer them books with a readability that matches their reading ability. It is also helpful to know the book's passage difficulty variation. If we want our students to learn to read by reading, then we want to give them material that fascinates, motivates, absorbs and also challenges them. We do that best by giving them books they want to read that are a little too hard for them, with passages that vary in passage difficulty. Then as they read along, they speed up and slow down. The speed-ups give them the energy and confidence needed to work through the slow-downs.

Using Lexiles

Figure 15
The Same Scale Measures
Book Readability & Reader Reading Ability



Books are brought into the Lexile Framework by Lexiling the books. Tests are brought into the Framework by Lexiling their items and using these Lexile calibrations as the basis for estimating readers reading ability.

Figure 16
A 1000 Lexile Slice Test Item

You don't just establish a character once and let it go at that. Dominant impression, dominant attitude, dominant goal, all the rest — they must be brought forward over and over again; hammered home in scene after scene, so that the audience has no opportunity to forget them.

Use _____ for emphasis.

- A. humor
- B. lighting
- C. repetition
- D. volume

To write a Lexile test item, we can use any natural piece of text. If we wish to write an item at 1000 Lexiles, we select books that contain passages at that level. We select a 1000 Lexile passage and add a relevant continuation sentence at the end with a crucial word missing. This is the "response illustration." Then we compose four one word completions, all of which fit the sentence but only one of which makes sense. Thus, the only technical problem is to make sure all choices complete a perfectly good sentence, but that only one choice fits the passage. The correct answer for the response illustration in Figure 2 is "Use repetition for emphasis."

The aim of a Lexile item is to find out whether the student can read the passage well enough to complete the response illustrated sentence with the word that fits the passage. Lexiled items like this are available at the Lexile website, www.lexile.com. Anyone can use them — any time.

The Lexile Slice is a simple easy to write item type. But in practice, we may not even need the slice to determine how well a person reads. Instead, we may proceed as we do when we take a child's temperature. Since, the Lexile Framework provides a ruler that measures readers and books on the same scale, we can estimate any person's reading ability by learning the Lexile level of the books they enjoy.

The One Minute Self-Report

When our child says "I feel hot!" we infer they have a fever. When a person says "I like these books," and we know the books' Lexile levels, we can infer that the person reads at least that well.

The Three Minute Observation

To find out more about our child, we feel their forehead. The three minute way to measure a person's reading is to pick a book with a known Lexile level and ask the person to "Read me a page." If they read without hesitation, we know they read at least that well. If they stumble, we pick an easier book. With two or three choices, we can locate the Lexile level at which the person is competent, just by having them read a few pages out loud. With a workbook of Lexile calibrated passages, we can implement the three minute observation this simply by opening the work book and turning the pages to give them successive passages to read.

The Fifteen Minute Measurement

To find out more, we use a thermometer to take our child's temperature, perhaps several times.

Figure 17
Taking a Measure

Method	Temperature Reading	
One Minute Self-Report	I have a fever!	I like this book!
Three Minute Observation	You feel hot!	Read this page.
Fifteen Minute Measurement	Your temperature is...	Your Lexile is...

For reading, we give the person some Lexiled passages ended with an incomplete sentence. To measure their reading ability, we find the level of Lexiled passages at which that person correctly recognizes what words are needed to replace the missing words 75 percent of the time.

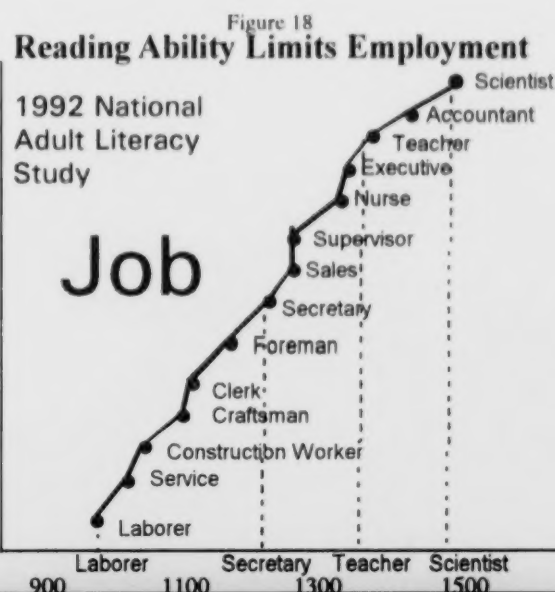
The Lexile reading ruler connects reading, writing, speaking, listening with books, manuals, memos and instructions. This stable network of reproducible connections empowers a world of opportunities of the kind that the inch makes available to scientists, architects, carpenters and tailors (Luce & Tukey, 1964).

In school, we can measure which teaching method works best and manage our reading curriculae more efficiently and easily. In business, we can Lexile job materials and use the results to make sure that job and employee match. When a candidate applies for a position, we can know ahead of time what level of reading ability is needed for the job and evaluate the applicant's reading ability by finding out what books they are reading and asking them to read a few sentences of job text out loud. This quick evaluation of an applicant's reading ability will show us whether the applicant is up to the job. When an applicant is not ready, we can counsel them, "You read at 800 Lexiles. The job you want requires 1000 Lexiles. To succeed at the job you want, you need to improve your reading 200 Lexiles. When you get your reading ability up to 1000, come back so that we can reconsider your application."

Lexile Perspectives

Job

Twenty-five thousand adults reported their jobs to the 1992 National Adult Literacy Study (Campbell et al, 1992; Kirsch, et al



1993, 1994). Their reading ability was also measured. Figure 18 summarizes the relationship between reading ability and employment. In 1992, the average laborer read at 1000 Lexiles. The average secretary at 1200. The average teacher at 1400. The average scientist at 1500.

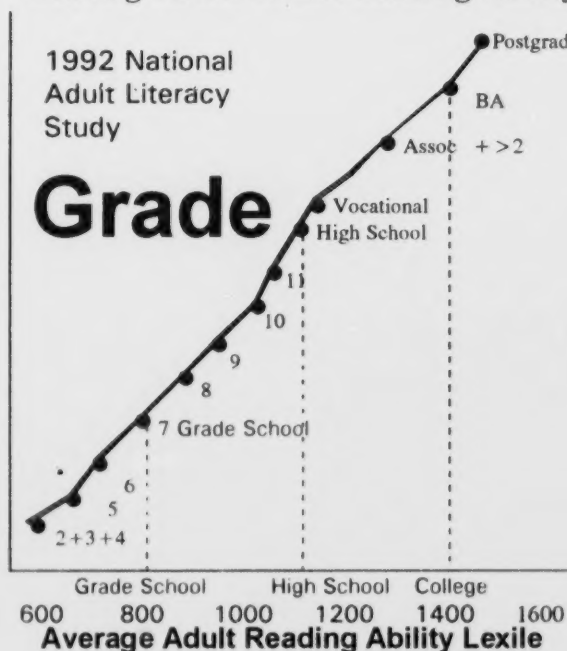
When we can see so easily how much increasing our reading ability can improve our lives, we cannot help but be motivated to improve, especially when what we must do is so obvious. If we want to be a teacher at 1400 Lexiles but read at only 1000, it is clear that we have 400 Lexiles to grow to reach our goal. If we are serious about teaching, the Lexile Framework shows us exactly what to do. As soon as we can take 1400 Lexile books off the shelf and read them easily, we know we can read well enough to be a teacher. But if we find that we are still at 1000 Lexiles, then we cannot avoid the fact that we are not ready to qualify for teaching, not yet, not until we teach ourselves how to read more difficult text.

School

Reading is learned in school. The 1992 National Adult Reading Study shows that there is a strong relationship between the last school grade completed and subsequent adult reading ability. Figure 19 shows that, on average, we are never more literate than the day we left school. The average 7th grade graduate reads at 800 Lexiles. The average high school graduate reads at 1150 Lexiles. College graduates can reach 1400 Lexiles. For many

of us, the last grade of school we successfully complete defines our reading ability for the rest of our lives. Once we leave school — and we no longer benefit from the reading

Figure 19
Leaving School Limits Reading Ability



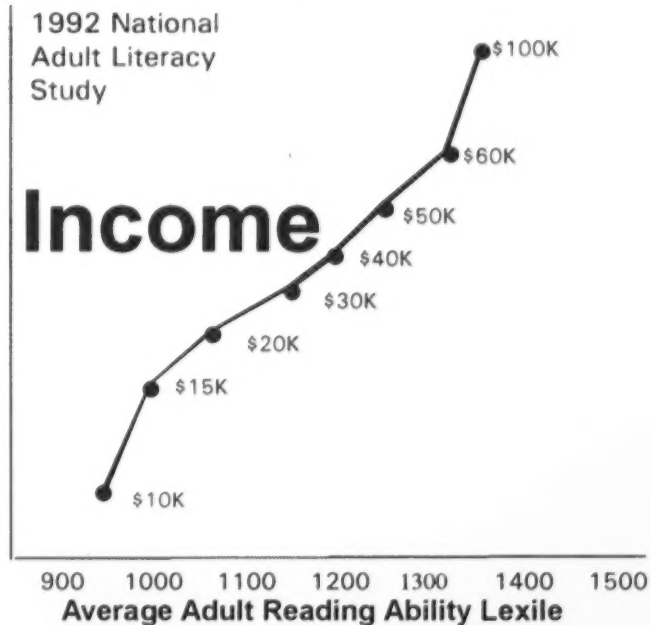
challenges that school provides — we tend to stop learning. The overwhelming implication of Figure 2 is that, if we aspire to become a truly literate society, then we must maintain schooling for everyone and help everyone stay in school, as long as possible.

Income

Reading ability also limits how much we can expect to earn. Figure 20 shows the average incomes of readers in the 1992 National Adult Literacy Study at various Lexile reading abilities. From 1000 to 1300 Lexiles, each reading ability increase of 150 Lexiles doubles our earning expectations. If we read at 1000 Lexiles and want to double our

potential, then we have to improve our reading to 1150 Lexiles. When students can see the financial consequences of reading ability on an easy to understand scale that connects reading ability and income, then they have a persuasive reason to spend more time improving their reading abilities. The simple relationship in Figure 20 makes the road to riches obvious and explicit. No need to berate students, "Do your home work!" Instead, we can show them, "You want more money? You want to be a doctor? Here is the road. Learn to read better. It's up to you. But we'll help you learn."

Figure 20
Reading Ability Limits Income



Reading Education

Education can only succeed if we connect learning to each learner's selfish motives. We need to involve our students individually, to engage their desires and arouse their drives. When we do that, student education will drive itself. Then, all we need do is to add support and guidance. Otherwise, we will continue to deceive ourselves into running a penitentiary system that keeps some troublesome kids off the street, but only for a while.

Remember, when we know text readability, all we need do to learn how well a student reads is to ask them to read a page or two aloud. If they succeed, we can give them a harder page. If not, we know their reading ability is below the readability of the text we asked them to read. No need for debate. No need for guesswork. No need for confusion or reproach. The student's status is plain to us and plain to them. We have not tricked them with a mysterious test score. All we have done is to help them see for themselves how high they can read.

Sources

- Bashaw, W.L. & Rentz, R.R. (1975). Equating Reading Tests with the Rasch Model, v1: Final Report & voll & vol2: Technical Reference Tables. *Final Report of U.S. Department of Health, Education, and Welfare Contract OEC-O72-5237*. Athens, GA: The University of Georgia. (ERIC Document Reproduction Nos. ED 127 330 & ED 127 331).
- Bashaw, W.L. & Rentz, R.R. (1977). The National Reference Scale for Reading: An Application of the Rasch Model. *Journal of Educational Measurement*, 14:161-179.
- Bernstein, I.H., & Teng, G. (1989). Factoring Items and Factoring Scales are Different: Spurious Evidence for Multidimensionality Due to Item Categorization. *Psychological Bulletin*, 105(3):467-477.
- Bormuth, J.R. (1966). Readability: New Approach. *Reading Research Quarterly*, 7:79-132.
- Campbell, A., Kirsch, I.S., & Kolstad, (1992). A. *Assessing Literacy: The Framework for the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Carroll, J.B., Davies, P. & Richmond, B. (1971). *The Word Frequency Book*, Boston: Houghton Mifflin.
- Chall, J.S. (1988). "The Beginning Years." In B.L. Zakaluk and S.J. Samuels (Eds.), *Readability: Its Past, Present and Future*, Newark, DE: International Reading Association.
- Crawford, W.J., King, C.E., Brophy, J.E. & Evertson, C.M. (1975). "Error Rates and Question Difficulty Related to Elementary Children's Learning," paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.
- Davis, Fred. (1944). Fundamental Factors of Comprehension in Reading, *Psychometrika*, 9:185-197.
- Educational Testing Service. (1990). *ETS Tests of Applied Literacy Skills*. NY: Simon & Schuster Workplace Resources.
- Jaeger, R.M. (1973). "The National Test Equating Study in Reading (The Anchor Test Study)." *Measurement in Education*, 4:1-8.
- Hitch, G.J., & Baddeley, A.D. (1974). Verbal Reasoning and Working Memory. *Journal of Experimental Psychiatry*, 28:603-621.
- Klare, G.R. (1963) *The Measurement of Readability*. Ames, IA. Iowa State University Press.

- Klare, G.R. (1974). "Assessing Readability." *Reading Research Quarterly*, 1:63-102.
- Kirsch, I.S., Jungeblut, A. & Campbell, A. (1991). *The ETS Tests of Applied Literacy*. Princeton, MJ: Educational Testing Service.
- Kirsch, I.S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult Literacy in America: A First look at the Results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics, U.S. Department of Education.
- Kirsch, I.S., Jungeblut, A., & Mosenthal, P.B. (1994). "Moving Toward the Measurement of Adult Literacy," paper presented at the March NCES Meeting, Washington, DC, 1994.
- Liberman, I.Y., Mann, V.A., Shankweiler, D. and Werfelman, M. (1982). "Children's Memory for Recurring Linguistic and Non-linguistic Material in Relation to Reading Ability." *Cortex*, 18:367-375.
- Lorge, I. (1939). "Predicting Reading Difficulty of Selections for School Children." *Elementary English Review*, 16:229-233.
- Loret, P.G.; Seder, A.; Bianchini, J.C. & Vale, C.A. (1974). *Anchor Test Study Final Report: Project Report* and vols. 1-30. Berkeley, CA: Educational Testing Service. (ERIC Document Nos. Ed 092 601 - ED 092 631.
- Luce, R.D. & Tukey, J.W. (1964). "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology*, 1:1-27.
- Miller, G.A. and Gildea, P.M. (1987). "How Children Learn Words." *Scientific American*, 257:94-99.
- Mitchell, J.V. (1985). *The Ninth Mental Measurements Yearbook*. Lincoln, NE: University of Nebraska Press.
- Rasch, G.A. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: The University of Chicago Press (first published in 1960).
- Reder, Stephen, (1996). "Dimensionality and Construct Validity of the NALS Assessment," in M.C. Smith (Ed.) *Literacy for the 21st Century: Research, Policy and Practice*, Greenwood Publishing, in Press.
- Rentz, R.R. & Bashaw, (1975). W.L. Equating Reading Tests with the Rasch Model, v1: *Final Report & vol1 & vol2: Technical Reference Tables. Final Report of U.S. Department of Health, Education, and Welfare Contract OEC-O72-5237*. Athens, GA: The University of Georgia. (ERIC Document Reproduction Nos. ED 127 330 & ED 127 331.
- Rentz, R.R. & Bashaw, W.L. (1977). The National Reference Scale for Reading: An Application of the Rasch Model. *Journal of Educational Measurement*, 14:161-179.

- Rinsland, H.D. (1945). *A Basic Vocabulary of Elementary School Children*.
- Rock, D.A., & Yamamoto, K. (1994). *Construct Validity of the Adult Literacy Subscales*. Princeton, NJ: Educational Testing Service.
- Salganik, L.H., & Tal, J. (1989). A Review and Reanalysis of the ETS/NAEP Young Adult Literacy Survey. Washington, DC: Pelavin Associates.
- Squires, D.A., Huitt, W.G., & Segars, J.K. (1983). *Effective Schools and Classrooms*, Alexandria, VA: Association for Supervisor and Curricular Development.
- Stenner, A.J., Smith, M., & Burdick, D.S. (1987). Fit of the Lexile Theory to Item Difficulties on Fourteen Standardized Reading Comprehension Tests. Durham, NC: Metametrics.
- Stenner, A.J. (1992). Meaning and Method in Reading Comprehension, paper presented at AERA, Division, D. Rasch Special Interest Group, S.F. CA.
- Stenner, A.J. (1996). Measuring Reading Comprehension with the Lexile Framework, Paper presented at the Fourth North American Conference on Adolescent/Adult Literacy, Washington, D.C., Feb.
- Stenner, A.J., & Burdick, D.S. (1997). "The Objective Measurement of Reading Comprehension: In Response to Technical Questions Raised by the California Department of Education Technical Study Group." Durham, NC: Metametrics.
- Stenner, A.J. "Writing Lexile Items" (1998). Paper presented at the American Educational Research Association Annual Meeting, University of Chicago, June.
- Thorndike, E. L. & Lorge, I. (1952). The Teacher's Word Book of 30,000 Word.
- Thorndike, E.L. & Hagen, E.P. (1965). *Measurement and Evaluation in Psychology and Education*, New York: Wiley.
- Thurstone, L.L. (1946). "Note on a Reanalysis of Davis' Reading Tests," *Psychometrika*, v11, n2, 185ff.
- Wright, B.D., & Stone, M. H. (1979) *Best Test Design*. Chicago: MESA Press.
- Woodcock, R.W., Woodcock (1974). *Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.
- Zeno, S.M., Ivens, S.H., Millard, R.T. & Davvuri, Raj. (1995). *The Educators Word Frequency Guide*, Touchstone.
- Zakaluk and S.J. Samuels (Eds.) (1988). *Readability: Its Past, Present and Future*, Newark, DE: International Reading Association.

Zwick, R. (1987). Assessing the Dimensionality of the NAEP Reading Data, *Journal of Educational Measurement*, 24:293-308.